Datafication & Discrimination

Koen Leurs (Utrecht University) & Tamara Shepherd (University of Calgary)

Revised version submitted 13 May 2016 for consideration in
SOCIAL RESEARCH IN THE AGE OF DATA. THE DATAFIED SOCIETY
Edited by Mirko Tobias Schäfer and Karin van Es
Amsterdam University Press

Introduction: Why datafication & discrimination?

Popular accounts of datafied ways of knowing implied in the ascendance of Big Data posit that the increasingly massive volume of information collected immanently to digital technologies affords new means of understanding complex social processes. The development of novel insights is attributed precisely to Big Data's unprecedented scale, a scale that enables what Viktor Mayer-Schönberger and Kenneth Cukier note is a shift away from causal inferences to modes of analysis based rather on "the benefits of correlation" (2013, p. 18). Indicating the vast implications of this shift, Mayer-Schönberger and Cukier's influential framing of Big Data describes a revolutionary change in the ways "we live, work and think," as phrased by the book's subtitle. But the "we" in this proclamation tends to go unspecified. Who exactly benefits from a shift toward correlative data analysis techniques in an age of Big Data? And by corollary, who suffers?

Our claim is that Big Data, given its origins in a western military-industrial context for the development of technology and concomitant mobilization within asymmetrical power structures, inherently discriminates against already marginalized subjects. This point has been raised in a number of critiques of the Big Data moment, for example in danah boyd and Kate Crawford's (2012) cautionary account of the mythologies of Big Data that obscure the ways it engenders new divides in data access, interpretation, representation, and ethics. Frank Pasquale (2015) has further illustrated the perils of "runaway data" that asymmetrically order our social and financial institutions through hidden algorithmic practices that tend to further entrench inequality by seeking to predict risk. An overview of the social inequalities perpetuated across various applications of Big Data can be found in the Open Technology Institute's series of primers on data and discrimination (Gangadharan, 2014). And yet, as the present collection attests, there may be ways of approaching Big Data with a critical lens in order for researchers to also benefit from new methods (see also Elmer et al., 2015). In the present collection, many authors make efforts to trouble the politics of Big Data by admonishing researchers to take alternative perspectives on data-based methods for social research: Nick Couldry, in line with his previous work on media's ontological implications (e.g., 2012), asserts that researchers should strive to "de-reify" social processes (p. 20); Carolin Gerlitz describes measurement as valuation that needs to be "de-naturalized" (p. 22); and Lev Manovich develops Cultural Analytics to get past demographic generalizations through a process of "estrangement" that prompts researchers to question their cultural assumptions (pp. 64, 67). Such actions – de-reifying, de-naturalizing, estranging – seem to offer important directives for highlighting the often invisible discriminatory functions of

Big Data, but how exactly do they work? Who and where are the actual people implicated in Big Data discrimination?

These kinds of questions seek to elaborate on Evgeny Morozov's contention, also in the present collection, that "social biases exist regardless of whether algorithms or computers are doing the job," and thus, "plenty of discrimination happens with regard to race, class, gender and so forth" (p. 25). One concrete example of this sort of discrimination is offered by Richard Rogers's explanation of query design, which describes how keyword searches illuminate the discursive operations of language within power dynamics (p. 80). His chief examples of keywords being used in this way, as part of "efforts at neutrality" between politically charged actors, include the BBC's use the term 'barrier' rather than 'security fence' or 'apartheid wall' to describe the Israeli-Palestinian border structure, and the preference for using the term 'conflict' diamonds or minerals above 'blood' diamonds or minerals on the part of industries attempting to inhabit a corporate social responsibility. These examples usefully point toward modes of political obfuscation that lie at the heart of data sets, as they are constituted from what are seen as legitimate sources of information to query. Even in these cases, however, what is getting queried are certain privileged accounts of political struggle, such as BBC coverage or industry discourse. Clearly, determining what counts in the first place as a legitimate object for Big Data analysis is a process that implicates deep-seated social biases at a number of levels, not only on the part of programmers and researchers but more fundamentally at the level of the organization of knowledge.

As a rejoinder to existing modes of talking about Big Data and what it means for social research, this chapter suggests an epistemological intervention from a critical, anti-oppressive stance that seeks to reinstate people within datafied social life. Rather than taking as its premise that Big Data can offer insights into social processes, this approach starts from the perspective of the people caught up in programs of social sorting, carried out by computational algorithms, particularly as they occupy marginalized positions within regimes of power-knowledge (to use Foucault's term). As a specifically situated case study, we examine the ways data are mobilized in European border control and how this phenomenon can be studied, framed through the eurocentric legacies of population measurement in colonial disciplinary surveillance. The connection between power and knowledge here is meant to implore researchers to consider how their deployments of Big Data, even from critical perspectives, may serve to replicate structures of discrimination by denying less "data-ready" ways of knowing. To that end, the conclusion of the chapter suggests some alternative methodological avenues for reinstating people – specifying who the "we" permits – in light of Big Data supremacy.

Datafied migration management and border control

Consider the following anecdote, placed here to illustrate a contemporary case example of the discriminatory workings of algorithmic sorting that separates the privileged from the unprivileged. Flashback to spring 2001. The first author, Koen, found himself sitting in an office cubicle on the 7th floor of a leading mobile phone provider in the Netherlands. Working as an activation and security & compliance officer in a life before academia, every once in a while he would get requests from local mobile phone shop

managers to look into activation requests which were 'rejected': "Why can't you activate this mobile phone contract?" Upon receiving a fax with signed contracts, passport copies, and bank statements, Koen and his fellow team of about 15 would assess applications in two ways. First, they would manually check for the applicant's financial history in the national credit registry. Second, the application would undergo an automated algorithmic risk assessment. The second process was opaque because the employees did not know exactly which variables were evaluated. Frequently, applicants who had a clean credit report were denied a contract after the algorithmic risk assessment. Over time, Koen discerned a pattern: those denied a contract were usually young men of non-Dutch descent who held temporary resident permits rather than Dutch passports and who were living in certain low-income areas. After asking his floor manager about this process, he received confirmation that the mal-payment prediction system targeted specific subgroups: "especially those Somalis, they never pay."

Many years later, algorithmic security assessments have become even more commonplace in the corporate world and have also gained prominence in various forms of state governance and surveillance. These systems are typically put in place to ensure greater predictive accuracy, according to a widespread faith in the insights generated through large data sets and statistical calculations since the development of statistical methods in the 1880s, when "the unflappable stance of quantitative method acquired a prestige that is still in force and whose power derives from the long valorization of impersonality" (Peters, 2001, p. 442). But what the anecdote illustrates is that, far from the imagined objectivity of social sorting through computation, people are still making the decisions at every step of the process (Gillespie, 2016).

Moreover, on a wider scale, the story of young male Somalis whose digital data traces algorithmically rendered them as undesired consumers finds parallels in the current moment of "refugee crisis" in Europe. Automated social sorting at state borders has become commonplace practice as part of governments' efforts to control flows of undesired migrants. For those privileged subjects carrying desirable passports, e-borders and iris scans sustain liquid flow across borders and planetary nomadic mobility as an effortless normality. By contrast, undesired subjects have to provide fingerprints – a genre of biometric data with a long history of criminal connotations – to be cross-referenced among a host of other identifiers in data-based risk calculations. Border control across the Global North and South is increasingly augmented by data collection and processing techniques developed by industry and ported to government applications. This digital policing of unwanted movement has been explored in previous studies on the Australian (Ajana, 2013) and Indian (Arora, 2016) contexts. In Europe, the division between desired and undesired migration plays out with its own specific contours, where although the internal Schengen Area is borderless, it controls against undesired external populations in the capacity of Fortress Europe.

Fortress Europe presents a particularly relevant context to study datafied discrimination because its contemporary practices of social sorting at the border show lingering traces of colonial-era human classification, measurement, and ordering, which were pioneered and mastered on subject populations in its peripheral territories throughout the last centuries. In recent years, continental Europe has prided itself on the premise of "Unity in Diversity" but as a "postcolonial location," it operates at its centre according to a

mostly hidden logic of "European apartheid" (Balibar, 2004, p. 121; Ponzanesi, 2016). Also spurred by the IS attacks in Paris in November 2015 and Brussels in March 2016, once again it mobilizes the colonial "idea of European identity as a superior one in comparison with all the non-European peoples and cultures" (Said, 2014, p. 7). Through various discursive, symbolic, material, and datafied processes, it decides who rightfully belongs to Europe by distinguishing between the west and the rest, or "the Orient" and the "Occident" (Said, 2014). Through this process the EU justifies who it retains, detains, or relocates, thereby distinguishing between lives worth living and "bare life": those non-citizens stripped of status who become unprotected by the law (Agamben, 2005).

Residual colonialism can be located in contemporary border policing. EU member states together manage nearly two thousand official entry-ports and 60 000 kilometres of land and sea borders (Broeders & Dijstelbloem, 2016, p. 247). In 2015, Europe welcomed a record of 611 million international tourist arrivals across these borders (UNWTO, 2016), while it sought to control over one million refugees who reached Europe across the Mediterranean sea, half of whom were fleeing war in Syria (UNHCR, 2015). This human sorting process of differentiating desired from undesired migrants is increasingly datafied through proliferating, non-linear, and non-geographically-bound electronic border governance processes, as a result of economic incentives, opportunistic political motives and the expansion of the security industry. Alongside the other centralized databases of the Schengen Information System and the Visa Information System, the European Dactyloscopy (EURODAC) biometric database, which holds fingerprints of asylum seekers and so-called "third-country nationals," is the most prominent datafied border control mechanism against unwanted others. EU counterterrorism measures have developed toward increased securitization and an outspoken desire to connect "data fragments" on non-Europeans through achieving higher "interoperability" between these and other architectures and databases (European Commission, 2016, p.3).

According to EURODAC regulations, "individuals from 14 years on should be fingerprinted, whether they are asylum seekers, aliens apprehended in relation with the irregular crossing of an external border or aliens found illegally present in a Member State" (EDPS, 2009, p. 15). 28 EU countries and four associated states have access to the fingerprint data in EURODAC. EURODAC is managed by the European Agency for the operational management of large-scale IT systems in the area of freedom, security and justice. Its data are stored at the headquarters in Talinn (Estonia), its operational management is housed in Strasbourg (France), and backup systems are in place in Sankt Johann im Pongau (Austria). In EURODAC jargon, "asylum applicants" and "aliens" become "data subjects" as their fingerprints are processed through three categorizations: 1) data of asylum applications; 2) data of "aliens" apprehended in connection with irregular border crossings; and 3) data related to "aliens" "illegally present" in member states (euLISA, 2015, pp. 13-14). Through algorithmic social sorting, the "Automated Fingerprint Identification System" classifies some data subjects as undesired, resulting in "digital deportability" (Tsianos & Kuster, 2013). Digitally sanctioned deportation may happen when a "hit" occurs: when a searched individual appears in the database, cross-referenced between the categories listed above.

The Central Database held over two million entries as of 2012 (Jones, 2014, p. 3), but evidencing the growing stream of Syrian refugees, it processed a total of 750 thousand

transaction requests in 2014, an 84% increase compared to 2012. Observing migration management in practice in Germany, researchers witnessed how the EURODAC system automatically played a James Bond melody whenever it "produced a hit" (Tsianos & Kuster, 2013, p. 10). This example of smart border gamification blurs boundaries between fact and fiction and exemplifies risks of dehumanization and depersonalization inherent to datafied social sorting. Individual people, faces, stories, and motives are not of interest to "smart" border processes. Furthermore, this echo of the popular culture version of espionage contrasts the automated, computational process of categorizing data subjects as undesirable, where the political struggle inherent in classification becomes naturalized through its bureaucratic manifestation (Bowker & Star, 1999, p. 196). Seemingly without human intervention, EURODAC operates as a disciplinary truth machine, making data-driven decisions. In 2014, 24% of EURODAC entries produced a hit proving a "data subject applied for asylum on two or more occasions", which means 121,358 people could be internally deported to a European member state where their fingerprints were previously taken. For 72,120 hits, data confirmed a data subject's "illegal" presence in Europe (euLISA 2015, p. 4, 15), which rationalizes and normalizes deportation to countries of origin, where individual deportees may experience persecution, torture or worse (Bloch & Schuster, 2005, pp. 496-497).

The United Nations High Commissioner for Refugees (UNHCR) has rightly criticized EURODAC for breaching the human right to privacy and family life – placing data subjects at significant risk, as data may be shared with countries of origin – and for stigmatizing groups, as fingerprints are associated with criminal activity and can result in latent fingerprint errors that cannot be eliminated (UNHCR, 2012). Although commonly presented as the perfect migrant management solution, such errors abound. In 2014, for example, from one million data entries, over one hundred thousand were rejected, largely due to data validation issues, fingerprint errors, or insufficient data quality (euLISA 2015, p. 18). These errors are typically attributed to tactics used to contest machine readability, including migrants' attempts to purposefully damage their fingertips with glue for this purpose (European Commission, 2015, p. 5). However, aging and heavy manual labour such as farming and construction work can also wear out fingerprints, causing insufficient image quality (Storisteanu et al 2015, p. 137; see also Tsianos & Kuster, 2013, pp. 33-35). Despite these problems, the EU has since made fingerprinting mandatory in response to the trend in 2014-2015 for Syrians and Eritreans refusing to be fingerprinted: "In cases where a EURODAC data-subject does not initially cooperate in the process of being fingerprinted […] it is suggested that all reasonable and proportionate steps should be taken to compel such cooperation," including steps like detention and coercion by force (European Commission, 2015, pp. 4-5). And fingerprints are only one subset of potential biometric data collection. If EURODAC officials encounter difficulties establishing the age of data subjects as over 14, they corroborate fingerprints with medical examinations that can include visual, dental, bone X-rays, blood tests, and sexual development tests, cross-referenced with connected datasets including census records.

The actions of EURODAC show how data subjects come to be constituted through a mixture of invasive and institutional strategies. The agency's own use of the term "data subjects" implies that people have a say over their data is compiled across diverse sources. And in fact, according to article 18(2) of the EURODAC regulation, migrants as

data subjects have the right to access their data. But there is a huge gap between theory and practice. In 2014, data subjects lodged only 26 such requests, a number that has decreased from 49 in 2013 and 111 in 2012 (euLISA, 2015, p. 18). These findings show that concerns previously voiced by the European Data Protection Supervisor (EDPS) over the lack of information available about consequences of being fingerprinted, the transmission of data, and rights of access, rectification, and deletion have not been addressed and so "the information provided to data subjects should be highly improved" (EDPS 2009, p. 14). But rather than data offering such subjectivity, datafied migration management evidences how migrants are subjected by data, with digital immobility and deportability – a dehumanized form of "exclusion through registration" (Broeders, 2011, p. 59) – as plausible social sorting outcomes.

Alongside database-led migrant management, the EU's "smart" border control processes demand scrutiny. Europe's securitization of its external land, sea, and aerial borders has resulted in the establishment of FRONTEX, the pan-European border agency, in 2004. In addition, in 2013, the European Border Surveillance System (Eurosur) was established to support the exchange of information between agencies "for the purpose of detecting, preventing and combating illegal immigration and cross-border crime" at the "external borders" (FRONTEX, 2016a). The agency lists "intelligence," "risk analysis," and "situational awareness" among its key missions, alongside its claim to operate in line with the EU fundamental rights charter. Providing lucrative business opportunities for a conglomerate of technology and arms manufacturers, this datafied surveillance arsenal combines radars, offshore sensors, on-shore olfactory sensors such as "sniffer" and "snoopy" satellite tracking systems, border patrol robots such as "Talos" on land and "Uncoss" at sea. Dronification is the latest step in this process, and a "common pre-frontier intelligence picture" is established through unmanned aerial vehicles, Remote Piloted Aircraft Systems, and Optionally Piloted Aircraft (Frontex, 2016a).

In addition to internal migration management which operates mainly out of the public eye, FRONTEX also maintains detailed statistics on various border movements such as interceptions, and routinely uses infographics to make available to the public such data. These data visualizations further reify distinctions between insiders and outsiders, the occident and orient. Figure 1 is an exemplary FRONTEX data visualization depicting continental Europe and representations of various flows of incoming migrants. The orange arrow on the far right visualizes that between January and March 2016, over 150 thousand people, mostly Syrian, Afghani, and Iraqi nationals, have attempted to illegally cross the border, taking the Eastern Mediteranean route (FRONTEX, 2016b).

Figure 1. Migratory Routes Map (FRONTEX, 2016b)

These kinds of data visualizations should be questioned, given emerging representational conventions such as mapping that serve to imbue data with "objectivity" based on ideological notions of "transparency, scientific-ness and facticity" (Kennedy et al., 2016, p. 716). Accordingly, the numerical and symbolical politics of this visualization can be unpacked: the choice to represent people with arrowheads taps into the symbolisms of weaponry, threat, and massive contagion, while the proportional size of the arrows does not consistently reflect actual figures of interceptions (blue arrow equals yellow in size, but only visualizes 675 crossings). FRONTEX's provision of data visualizations is thus illustrative for its appeals to transparency and accountability but also for its ideological thrust. EU bordering is characterized by a paradoxical situation of painful "exclusion from registration" (Broeders, 2011, p. 59): a significant gap remains between what is recorded and what remains untracked. It is striking, for example, that the agency does not systematically gather data on deaths at the borders (MigrantFiles, 2015). Prioritizing one "smart"-border statistic over another reflects a poignant decision, given that Europe has recently been considered as the "deadliest migration destiny of the world," with the Mediterranean becoming an "open air cemetery" (S. Wolff, 2015). These crucial representational decisions lie at the heart of discriminatory data practices, which seem to maintain a longstanding appeal to quantitative objectivity despite the historical precedent for "people in the algorithm" (Gillespie, 2016).

From statistical subjects to datafied society

Contemporary cases around migration and border security offer a salient entry point into the discriminatory implications of a datafied society. But the current faith in Big Data as a font for accurate representations of and predictions about social groups has a much longer history. In terms of predicting risk, for example, the insurance industry was the site of key innovations that exemplify discriminatory practices at all levels of statistical calculation. Precedents for the cases of predictive discrimination in insurance redlining

examined by Pasquale might be found in Dan Bouk's (2015) history of the American insurance industry's expansion in the late nineteenth century. By this time, insurance companies sought to expand beyond their traditional clientele of middle class white men in the Northeastern states. Bouk charts the development of data-based metrics that asymmetrically created (not simply calculated) the higher risk factor of groups such as women, children, Southerners, and African-Americans, which served to not only inform differential insurance policies but also to simultaneously construct vast swathes of the US population as differentially valued "statistical subjects," the precursors to contemporary "data subjects."

One key story in this process was how Prudential's insurance policies were developed in the 1890s based on statistician Frederick Hoffman's measurements and calculations, including data gleaned from the tombstones of pre-Civil War segregated cemeteries and contemporaneous eugenic science (Bouk, 2015, pp. 113-120). Hoffman "created the largest compilation of data about the American Negro then available in print," in support of Prudential's categorization of Southern African-American clients as higher risk (M. J. Wolff, 2006, p. 85). While African-American activists mobilizing against this framing successfully pressured lawmakers to introduce anti-discrimination legislation, race-based discrimination in the insurance industry persists to this day:

> Race-based pricing classifications and coverage restrictions proved difficult to dislodge not only because of the structure and legal regulation of private commercial insurance markets, but also because of the strength of the underlying ideologies of racial difference, race separation, and the rhetorical power of actuarial language. Legislation and litigation, despite some progress, proved ineffective in changing industry practice. (Heen, 2009, p. 362)

As a historical precedent, the story of discriminatory insurance practices illustrates a number of crucial components of the modes of discrimination underlying popular beliefs about the power of Big Data today. Despite the seeming robustness of large data sets, as the European migration case illustrates, structural biases at every moment of "calculation" – data gathering, organization, aggregation, algorithmic design, interpretation, prediction, and visualization – serve to construct legitimized difference by reproducing existing inequalities across individuals as data subjects.

Despite their conceit to objectivity, data-based calculations reinforce inequalities specific to historical conjuncture. In the story of insurance redlining, it was post-Civil War race relations and the eugenics movement that informed data gathering and organization. In the story of today's "refugee crisis," datafication is shaped by EU economic policy along with racial and ethnic stereotypes dating back to the colonial era. Whatever the historical moment, social context is key for understanding the specific modes of inequality embedded in quantitative operations. In turn, the application of data-driven insights to perceived social problems – typically framed around mitigating risk – performs a doubling of discriminatory frameworks through the asymmetrical representation of particular social problems that constructs them precisely as "problems." Representations of contemporary European migration, as discussed above, draw on disproportionately datafied renderings of certain groups of refugees, replicating fears of the other that stem from the colonial era. These representations show how the social thus becomes an

effect of data as a resource to be appropriated (Couldry & Van Dijck, 2015). The cycle of discriminatory measurement being used to inform discriminatory representations is one important point of continuity between the early uses of statistics and the current fashion for Big Data.

Another salient point of historical comparison concerns the role of the military-industrial complex in leading the way with the development and implementation of data-based "solutions" to what are construed as social problems. Indeed, data-based discrimination as a military-industrial process seems not to be effectively contained by public governance or law; this goes for insurance as well as other sectors working to construct the statistical or data subject. The datafication of credit, medicine, marketing, human resources, policing, urban planning, transportation, and security industries over the course of the twentieth century laid the groundwork for the ascendance of Big Data according to developments in information processing (Danna & Gandy, 2002). The exponentially increasing capacity for collecting and tabulating social dynamics as information has been framed in James Beniger's (1989) classic text as a "control revolution" coalescing around Big Science in the 1950s and 60s. At this time in the US, funding for the development of tools designed to gather, process, and store increasingly large data sets was allocated according to a Cold War rationale intended to strengthen national security and military intelligence operations. The internet itself – the meta-network that supports proliferating data – of course emerges from within this context, as is evident in vestiges such data "caches" organized according to "C3I" protocols (command, control, communication and intelligence) which operate behind the screen (Ricker Schulte, 2015, p. 40; see also: Gitelman, 2006, p. 114; Norberg & O'Neill, 1996). It is this eurocentric and masculinist ideological nexus that informs the discriminatory considerations designed into data science technologies and techniques that, when combined with the commercialization imperatives of industry, form the basis for the networking of statistical subjects together within a datafied society.

What changes in the move from a statistical society to a datafied society is framed by proponents of Big Data as economies of scale. Innovative insights emerge from networking between unprecedented large data sets, which creates value far beyond any one set alone (Mayer-Schönberger & Cukier, 2013, p. 135). Yet despite these economies of scale, similar operations that have historically encoded discrimination into statistical calculations remain. The implication of networking large data sets is that at every level, from individual data subjects – and even more finely, "dividuals," or sub-individual units (Terranova, 2004, p. 34) – up to entrenched industries and institutions, epistemological and ideological contours around what counts and how it is measured still serve to produce and reinforce structural inequality. Unlike traditional statistical analysis, however, Big Data methods perform these operations in ways that are often automated and invisible. Quintessential examples of contemporary social life being reorganized according to discriminatory datafication include Facebook's profiling of users' "ethnic affinities" based on their online activities, Amazon's redlined same-day delivery zones, or Google's culturally and linguistically biased search results (Knobel & Bowker, 2011; Noble, 2016). The invisibility of the algorithmic biases underlying such social platforms enable the "laundering of past practices of discrimination" so that they become black-boxed, "immune from scrutiny" (Pasquale, 2015, p. 41).

Further, the development of social sorting algorithms in commercial contexts informs contemporary modes of governance, as seen in the case of migration – enabled by a permanent state of exception preserved through the "war on terrorism" (Guzik, 2009) – where police surreptitiously collect data and run predictive calculations that violate the privacy of people living in racialized communities (Crawford & Schultz, 2014), and the state distinguishes citizenship from "foreignness" according to the hidden logic of the National Security Agency's surveillance assemblage (Cheney-Lippold, 2016). The public implication of this shift toward opaque and automated datafied discrimination renders justice through transparency and accountability ever more elusive for data subjects (Barocas & Selbst, 2016).

The fact that datafication supports the increasing automation and opacity of discriminatory practices not only accords with what Beniger foresaw as the centrality of information processing for social control, but also points toward immanent surveillance as the crux of domination. The military-industrial invention of the datafied individual can be seen as more deeply embedded within a colonial legacy of surveillance as the means of achieving the dual purpose of value extraction and social control (Mbembe, 1992). While in a datafied society one might think instead of "dataveillance," the continuous monitoring that pervades the social fabric through immanent collection of both data and metadata "for unstated present purposes" (Van Dijck, 2014, p. 205; see also Zimmer, 2008), the ways algorithms exert discriminatory "soft bio-power" find precursors in surveillance as a disciplinary gaze (Cheney-Lippold, 2011).

In order to collect people's information, they must be observed, their pertinent information discerned, translated into a notation system, and organized. Each of these steps involves a surveillant gaze whose roots can be traced to military-industrial colonial expansion that relied on making use of indigenous populations through techniques based on visible measurement (Glissant, 1997). This profit-oriented domination was further justified through European imperial knowledge systems such as medically proven "inferiority" which rationalized the existence of "subject races" (e.g. non-whites) that needed to be ruled by white superiors as part of their civilizing mission (Said, 2014; Wekker, 2015). Disproportionate applications of surveillance to othered bodies characterized the slave trade as a network of power exercised through a monopoly of knowledge comprised of overseers, paper technologies, shipping routes, biometrics, plantations, identification cards, and the census (Browne, 2015; Siegert, 2006). Parallel to contemporary data visualization and the social graph, these various practices of surveillance generated data that could be processed, organized, and most importantly, mapped onto the expanding territories of empire (Shepherd, 2015). Moreover, the ways in which diverse data sources could be networked through mapping highlights the eurocentric conceit of datafication; as José Rabasa (1993, p. 180) contends, traces of European expansionism continue to imbue measurements and representations of the social world with an underlying eurocentric universality – this is the basis for the "we" who benefit from Big Data in Mayer-Schonberger & Cukier's account.

An epistemological critique thus subtends a look backward to the historical precedents for applications of data processing to social organization. The ways knowledge is produced through measurement invokes a culturally specific set of baggage inherent in the very language used to delimit what counts as "data," as the first step in a series of

human processes that seek to make sense of the social world through measurement and organization. Even within a Eurocentric lineage of structuralist and poststructuralist thought (e.g., Foucault's *Les mots et les choses*, 1966), human sciences are acknowledged to be governed by an unconscious set of formative rules as productive of knowledge as power. In this light, the work of nineteenth-century anthropometrist Adolphe Quetelet, who believed that data sets large enough would produce accurate predictions of criminality (Beirne, 1987), might be seen as a formative epistemological touchstone for the contemporary faith in the "bigness" of Big Data. Belief in quantification is a hallmark of this episteme, embedded in the language, tools, methods, and data itself used to represent social life. In other words, the software that immanently conducts operations of collection, organization, and prediction might be seen as the "frozen organizational and policy discourse" that circulates as a means of legitimizing inequality (Bowker & Star, 1999, p. 135).

Methodological interventions and alternative cartographies

Therefore the question arises whether we can repurpose big data approaches for anti-oppressive knowledge production, and, if so, how? We explore this question by offering some methodological considerations developed from the situated case of human mobility. Clearly, the methodological debate on Big Data is polarized. Utopianists celebrate big data as the next lucrative frontier (colonial metaphor intended). In line with this optimistic rhetoric, quantitative and mathematically oriented scholars have praised Big Data for natural, unmediated, objective, purer, and self-explanatory access to social processes. Exemplary of this discourse is the recognition of the "data scientist" as the "Sexiest job title of the 21st Century" industry (cited in Gehl, 2015). On the other end of the spectrum, a dystopian denouncement of Big Data as a form of "methodological genocide" highlights its lack of attention to history, culture, context, specificity, meanings, structure, and agency (Uprichard, 2015).

Realistically, because datafication is a pressing contemporary empirical geo-political reality, we cannot simply reject dealing with it, nor should we univocally champion it as a silver bullet. Evelyn Ruppert and her colleagues suggest that data-mining can be of value for socio-cultural research through "specific mobilizations" of certain digital methods "in particular locations" (2013, p. 32). Similarly, the chapter on reflexive research by Karin van Es, Nicolás López Coombs and Thomas Boeschoten in the present collection (pp. 101-106) offers a promising starting point for opening up onto more critical methodological approaches to the datafied society. However, to date, little attention has been given how to make data mining a people-centred process, which accounts for dynamism, complexity, reflexivity, diversity, and multiplicity (Leurs, forthcoming).

The common paradigm of disembodied, impartial knowledge production is often conducted on the basis of a utilitarian research ethics. In large-scale data-driven research projects, ethical safeguards are commonly geared toward managing risk and reputation on the part of the institution, rather than protecting those people involved in the study. This model draws on a cost-benefit analysis and abides by expectations held by bio-medical-ethically oriented university Institutional Review Boards. Consider for

example the painful example of the "massive-scale emotional contagion" experiment, where a team of academic and industry researchers manipulated the News Feeds of nearly seven hundred thousand unwitting Facebook users. They took Facebook's Terms of Service as a proxy for informed consent, and no-one could opt out from the study (Kramer, Guillory & Hancock, 2014).

By contrast, providing distinctive alternative ethical positions stemming from feminist, critical pedagogical, anti-oppressive, community-based, and indigenous paradigms and methods offer ways to rethink big data as a progressive toolkit. Although they have yet to engage in sustained dialogue with data studies, these approaches share a common interest in dialogically involving informants as knowledge co-producers or co-researchers who share valuable insights. They involve consciously highlighting the human in data-based decisionmaking, where each step in the process of creating and manipulating large datasets involves ethical reflexivity (O'Neil, 2016). With ambitions to decolonize dominant disembodied research methodologies that claim positivist objectivity, scholars in alternative paradigms seek to establish greater reciprocity for underprivileged, queer, indigenous, or otherwise non-mainstream voices (Kovach, 2010; Walter & Andersen, 2013). Rather than neutrally extracting self-explanatory data from an apolitical data void, scholars may account for power relations, and prioritize listening, relationalities, fluidity, journeying, mutual trust, and strategic refusals as a way of helping data subjects regain sovereignty over knowledge production. When transposed to the digital context, these approaches may prompt scholars as activists to take serious the agency of individuals over their own information; this data can be collaboratively repurposed for community advocacy (Gubrium & Harper, 2013).

Alternative cartographies and bottom-up initiatives that reinstate what's missing in Big Data are exemplary for how data-based initiatives can be appropriated for community advocacy, "civic action" (Schäfer, 2016), "agency" (Kennedy, Poell and Van Dijck, 2015), and "data activism" (Milan & Gutiérrez, 2015). Various digital counters and mapping initiatives maintained by consortia of journalists, researchers, and activists that combine big and small data (individual cases) have made a growing impact. For example, the Bureau of Investigative Journalism (BIJ) manage wherethedronesstrike.com, visualizing 10 years of drone strikes in Pakistan. The Missing Migrant Project maintained by the International Organization of Migration tracks "deaths along migratory routes worldwide." The Migrant Files similarly maps "the human and financial cost of 15 years of Fortress Europe," providing "data-driven insights on migration to Europe."

Figure 2. Abdul Satar's story

The BIJ's recent infographic, Which countries treat children like children? (Figure 2), provides access to voices and accounts of 95 thousand unaccompanied children seeking asylum in Europe in 2015. Besides numerical overviews, situated individual voices are integrated into the map. For example, we can find out about Abdul Satar's story: We see a young man in a photograph. This 17-year old Syrian currently lives in London. He fled from his native country in 2013. In a video, we hear how he narrates his journey in Arabic, with English subtitles. This journey is also visualized, alongside the complete translated transcript of the interview. His claim, "Someone must talk about us – because no one is listening to us," resonates through the small-data contextual cues juxtaposed against the larger data set.

Such initiatives commonly combine various sources of people-centred information, crowdsourced data gathering, and open access databases that accommodate non-specialist public audiences for the information. For further examples of how alternative data corpuses can draw attention for marginalized issues, consider this open access spreadsheet which lists cross-referenced deaths of migrants at the European borders since 2000, this archive provides geo-tagged data on migrants deaths between 2014-2016 globally, and this open access spreadsheet lists EU deportations from 2000. While combining large and small data sets and toggling between distant and close reading can risk a reification of the methods themselves (Caplan, 2016), the point to emphasize in alternative cartographies is the reflexive self-positioning of the observer. In this way, small data might be reconceptualized as "deep data," information rendered self-aware through cultural continuity that acknowledges data's formative epistemological contexts (Brock, 2015). These embodied, situated, and re-humanized examples of doing "deep data" analysis offer incentives to further think about ways in which Big Data might be strategically mobilized as an anti-oppressive knowledge-power system.

Conclusion: Who are "we" in the datafied society?

Returning to the proclamation that Big Data affords a revolutionary change in the ways "we live, work and think," as phrased by Mayer-Schönberger and Cukier's subtitle, the "we" might be most fruitfully interrogated from the perspective of power and privilege. Their popular account of Big Data may indeed be seen as a continuation of longstanding power structures within the mythologies of information processing. Echoing Cisco CEO John Chambers's claim in the 1990s that the internet would "change the way we work, live, play and learn" (Fryer & Stewart, 2008), the rise of Big Data extends Beniger's control revolution, which opens with the conceit that "understanding ourselves in our own particular moment in history will enable us to shape and guide that history" (1986, p. 6). If the goal of data-based inquiry is to shape and guide history, and if "methods are social practices, means of forming good communities, not just tools for poking at reality" (Peters, 2011, p. 444), then specifying the "we" who are invested with agency and "they" who become excluded others in a datafied society is critical. Data analytics dashboards seductively promise a complete rendering of reality, but any approach to data-based social research must contend with the thorny question: who are "we"? And how can "we" be critiqued, opened up, accessed, and delineated? As shown in the contemporary case of migration, situated within a colonial history of the statistical measurement of populations, data's inherent discriminatory operations need to be uncovered in order to get at the "we." This involves recognizing the politics embedded within technological artifacts (e.g., Winner, 1980), and particularly recognizing the asymmetrical power embedded in a doctrine of objectivity "honed to perfection in the history of science tied to militarism, capitalism, colonialism, and male supremacy" (Haraway, 1991, p. 187). Taking this specific assemblage of privilege into account, researchers working with data can counter the investments in its "bigness" with anti-oppressive tactics drawn from small and deep data co-construction in order to lay bare the ethical implications of a datafied society.

References

Agamben, G. (2005). *State of exception*. Chicago: University of Chicago Press.

Ajana, B. (2013). *Governing through biometrics: The biopolitics of identity.* Basingstoke: Palgrave Macmillan.

Arora, P. (2016). The bottom of the data pyramid: Big data and the global South. *International Journal of Communication, 10*, 1681-1699.

Balibar, E. (2004). *We, the people of Europe?*. Princeton: Princeton University Press.

Barocas, S., & Selbst, A. D. (2016, in press). Big data's disparate impact. *California Law Review 104*.

Beirne, P. (1987). Adolphe Quetelet and the origins of positivist criminology. *American Journal of Sociology, 92*(5), 1140-1169.

Beniger, J. (1989). *The control revolution: Technological and economic origins of the information society*. Cambridge, MA: Harvard University Press.

Bloch, A., Schuster, L. (2005). At the extremes of exclusion: Deportation, detention and dispersal, *Ethnic and Racial Studies*, 28(3), 491-512.

Bouk, D. (2015). *How our days became numbered: Risk and the rise of the statistical individual*. Chicago: University of Chicago Press.

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. Cambridge, MA: MIT press.

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society, 15*(5), 662-679.

Brock, A. (2015). Deeper data: a response to boyd and Crawford. *Media, Culture & Society, 37*(7), 1084-1088.

Broeders, D. (2011) "A European 'border' surveillance system under construction', in H. Dijstelbloem & A. Meijer (Eds.) *Migration and the new technological borders of Europe*. Basingstoke: Palgrave: 40-67.

Broeders, D., & Dijstelbloem, H. (2016). The datafication of mobility and migration management: The mediating state and its consequences. In I. Van der Ploeg and J. Pridmore (eds.), *Digitizing identities: Doing identity in a networked world* (pp. 242-260). London: Routledge.

Browne, S. (2015). *Dark matters: On the surveillance of blackness*. Durham, NC: Duke University Press.

Caplan, L. (2016). Method without methodology: Data and the digital humanities. *E-flux, 72*(4): http://www.e-flux.com/journal/method-without-methodology-data-and-the-digital-humanities/

Cheney-Lippold, J. (2016). Jus Algoritmi: How the national security agency remade citizenship. *International Journal of Communication, 10*, 1721-1742.

Cheney-Lippold, J. (2011). A new algorithmic identity: Soft biopolitics and the modulation of control. *Theory, Culture & Society, 28*(6), 164-181.

Couldry, N. (2012). *Media, society, world: Social theory and digital media practice*. London: Polity.

Couldry, N., & van Dijck, J. (2015). Researching social media as if the social mattered. *Social Media + Society, 1*(2): http://sms.sagepub.com/content/1/2/2056305115604174.full

Crawford, K., & Schultz, J. (2014). Big data and due process: Toward a framework to redress predictive privacy harms. *Boston College Law Review, 55*(1), 93-128.

Danna, A., & Gandy Jr, O. H. (2002). All that glitters is not gold: Digging beneath the surface of data mining. *Journal of Business Ethics, 40*(4), 373-386.

Elmer, G., Langlois, G., & Redden, J. (Eds.). (2015). *Compromised Data: From Social Media to Big Data*. New York: Bloomsbury Publishing USA.

EDPS (2009). Eurodac Supervision Coordination Group Second Inspection Report Eurodac Supervision Coordination Group Second Inspection Report: https://secure.edps.europa.eu/EDPSWEB/webdav/shared/Documents/Supervision/Eurodac/09-06-24_Eurodac_report2_EN.pdf

European Commission (2015). Commission staff working document on Implementation of the Eurodac Regulation as regards the obligation to take fingerprints: http://ec.europa.eu/dgs/home-affairs/e-library/documents/policies/asylum/general/docs/guidelines_on_the_implementation_of_eu_rules_on_the_obligation_to_take_fingerprints_en.pdf

European Commission (2016). Stronger and smarter information systems for borders and Security. Communication from the commission to the European Parliament

and the Council: http://www.eulisa.europa.eu/Newsroom/News/Documents/SB-EES/communication_on_stronger_and_smart_borders_20160406_en.pdf
euLISA (2015). Annual report on the 2014 activities of the Central System of Eurodac. http://www.eulisa.europa.eu/Publications/Reports/Eurodac%202014%20Annual%20Report.pdf
Foucault, M. (1966). *Les mots et les choses*. Paris: Gallimard.
FRONTEX. (2016a). Legal basis: http://frontex.europa.eu/about-frontex/legal-basis/
FRONTEX. (2016b). Migratory routes map: http://frontex.europa.eu/trends-and-routes/migratory-routes-map/
Fryer, B., & Stewart, T. (2008). Cisco sees the future: The HBR interview with John Chambers. *Harvard Business Review, 86*(11), 72-79.
Gangadharan, S. (Ed.). (2014). *Data and Discrimination: Selected Essays*. Washington, DC: Open Technology Institute, New America Foundation. https://www.newamerica.org/oti/data-and-discrimination/
Gehl, B. (2015). Sharing, knowledge management and big data: A partial genealogy of the data scientist. *European Journal of Cultural Studies*, 18(4-5), 413-428.
Glissant, E. (1997). *Poetics of Relation*. B. Wing (trans.). Ann Arbor: University of Michigan Press.
Gillespie, T. (2016). Facebook Trending: It's made of people!! *Culture Digitally*, 9 May: http://culturedigitally.org/2016/05/facebook-trending-its-made-of-people-but-we-should-have-already-known-that/
Gitelman, L. (2006). *Always already new: Media, history and the data of culture*. Cambridge, MA: MIT Press.
Gubrium, A. & Harper, K. (2013). *Participatory visual and digital methods*. Walnut Creek, CA: Left Coast Press.
Haraway, D. (1991). *Simians, cyborgs, and women: The reinvention of nature*. New York: Routledge.
Heen, M. L. (2009). Ending Jim Crow life insurance rates. *Northwestern Journal of Law & Social Policy*, *4*, 360-399.
Kennedy, H., Hill, R.L., Aiello, G. & Allen, W., 2016. The work that visualisation conventions do. *Information, Communication & Society,* 19(6), 715-735.
Kennedy, H., Poell, T., Van Dijck, J. (2015). Introduction: Special issue on Data and agency. *Data & Society*. December, DOI: 10.1177/2053951715621569.
Kovach, M. E. (2010) *Indigenous methodologies: Characteristics, conversations, and contexts*. Toronto: University of Toronto Press.
Knobel, C., & Bowker, G. C. (2011). Values in design. *Communications of the ACM, 54*(7), 26-28.
Kramer, A., Guillory, J., Hancock, J., (2014). Experimental evidence of massive-scale emotional contagion through social networks. *PNAS*, *111*(29), 8788-8790.
Leurs, K. (forthcoming 2017). Feminist and postcolonial data-analysis: using digital methods for ethical, reflexive and situated sociocultural research. Lessons learned from studying young Londoners' digital identities. *Feminist Review*, themed issue on 'Where are we with feminist methods?'
Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Boston: Houghton Mifflin Harcourt.
Migrantfiles. (2016). The human and financial cost of 15 years of Fortress Europe: http://www.themigrantsfiles.com/

Milan, S., & Gutiérrez, M. (2015). Medios ciudadanos y big data: La emergencia del activismo de datos. *MEDIACIONES,* (14), 10-26.

Mbembe, A. (1992). Provisional notes on the postcolony. *Africa: Journal of the International African Institute*, *62*(1), 3-37.

Noble, S. (2016). *Algorithms of oppression: Race, gender and power in the digital age*. New York: New York University Press.

Norberg, A. L., & O'Neill, J. E. (1996). *Transforming computer technology: Information processing for the Pentagon, 1962-1986*. Baltimore: Johns Hopkins University Press.

O'Neil, C. (2016). How to bring a better ethics to data science. *Slate*, 4 February: http://www.slate.com/articles/technology/future_tense/2016/02/how_to_bring_better_ethics_to_data_science.html

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge, MA: Harvard University Press.

Peters, J. D. (2001). "The only proper scale of representation": The politics of statistics and stories. *Political Communication, 18*(4), 433-449.

Ponzanesi, S. (2016). Connecting Europe: Postcolonial mediations. Inaugural lecture Utrecht University: www.uu.nl/file/43575/download?token=mZlBGt8R.

Rabasa, J. (1993). *Inventing America: Spanish historiography and the formation of Eurocentrism*. Norman, OK: University of Oklahoma Press.

Ricker Schulte, S. (2015). *Cached: Decoding the Internet in Global Popular Culture*. New York: NYU Press.

Said, E. (2014). *Orientalism*. *25th anniversary edition.* New York, NY: Vintage Books.

Schäfer, M. (2016). Introduction to special issue: Challenging citizenship: Social media and big data, *Computer Supported Cooperative Work*, 25(2), 111–113.

Shepherd, T. (2015). Mapped, measured, and mined: The social graph and colonial visuality. *Social Media + Society, 1*(1): http://sms.sagepub.com/content/1/1/2056305115578671.full

Siegert, B. (2006). *Passagiere und Papiere: Schreibakte auf der Schwelle zwischen Spanien und Amerika*. Munich: Fink.

Storisteanu, D., Norman, T., Grigore, A. & Norman, T. (2015). Biometric fingerprint system to enable rapid and accurate identification of beneficiaries. *Global Health* 3(1), 135-137.

Terranova, T. (2004). *Network Culture: Politics for the Information Age*. London: Pluto Press.

Tsianos, V. & Kuster, B. (2013). Thematic report 'Border Crossings'. *MIG@NET. Transnational digital networks, migration and gender:* http://www.mignetproject.eu/?p=577

Uprichard, E., (2015). Most big data is social data - the analytics need serious interrogation. Philosophy of Data Science. *Impact of Social Science Blog*, London School of Economics and Political Science. Available at: http://blogs.lse.ac.uk/impactofsocialsciences/2015/02/12/philosophy-of-data-science-emma-uprichard/

Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society, 12*(2), 197-208.

Walter, M., & Andersen, C. (2013). *Indigenous Statistics: A Quantitative Research Methodology*. Walnut Creek, CA: Left Coast Press.

Walters, W. (2005). *Rethinking Borders Beyond the State*. Basingstoke: Palgrave.

Wekker, G. (2015). *White Innocence.* Durham, NC: Duke University Press.
Winner, L. (1980). Do artifacts have politics? *Daedalus, 109*(1), 121-136.
Wolff, S. (2015). Deaths at the border. Scant hope for the future. *Clingendael:* http://www.clingendael.nl/publication/deaths-sea-scant-hope-future?lang=nl
Wolff, M. J. (2006). The myth of the actuary: Life insurance and Frederick L. Hoffman's "Race Traits and Tendencies of the American Negro." *Public Health Reports, 121*(1), 84-91.
Zimmer, M. (2008). The gaze of the perfect search engine: Google as an infrastructure of dataveillance. In A. Spink & M. Zimmer (eds.), *Web Search* (pp. 77-99). Berlin: Springer.